Modeling Route Choice Behavior From Smartphone GPS data*

Michel Bierlaire † Jingmin Chen † Jeffrey Newman †

October 16, 2010

Report TRANSP-OR 101016
Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne
transp-or.epfl.ch

*This research is supported by the Swiss National Science Foundation grant 200021/131998 - Route choice models and smart phone data

†Transport and Mobility Laboratory, Ecole Polytechnique Fédérale de Lausanne, CH-1015 Lausanne, Switzerland

Abstract

Smartphones have the capability of recording various kinds of data from built-in sensors such as GPS in a non-intrusive, systematic way. In order to be used as observations for route choice models, the discrete sequences of GPS data need to be associated with the transportation network to generate meaningful paths. In this paper, a probabilistic path generation algorithm is proposed to replace conventional map matching (MM) algorithms. Instead of giving a unique matching result, the proposed algorithm generates a set of potential true paths. Temporal information (speed and time) is used to calculate the likelihood of the data while traveling on a given path. Comparisons against a state of the art deterministic MM algorithm using real trips recorded from a single user's smartphone are performed so as to illustrate the robustness and effectiveness of the proposed algorithm. Also, a Path-Size Logit (PSL) model is estimated based on a sample of real observations. The estimation results show the viability of applying the proposed method in a real context.

Keywords: route choice modeling, GPS data, path observation generation, map matching, network-free data

1 Introduction

Developing technology has long been harnessed to supplement or replace parts of travel behavior surveys. Tools such as GPS tracking devices have been used to track movements of individuals in a systematic way, instead of relying merely on travel diaries and prompted recall questioning. Tracking survey participants using a specialized GPS device provides some challenges. In particular, people may forget to charge the device, or leave it at home. Nowadays, many people carry a wireless phone. They already manage the tasks of charging and remembering to carry it, at least as well as for any special survey device. Therefore, we propose, as in Stopher [2008], to bundle the survey data collection into a phone.

An important feature of most GPS capable cell phones is Assisted-GPS, which reduces warm-up time for getting the first GPS reading to seconds. This advantage provides more opportunities to observe full tracks of the user's trips without losing the beginning parts of trips. However, the GPS device consumes a great deal of energy. Due to practical constraints, such as limited phone storage space and expensive data transmission cost, data

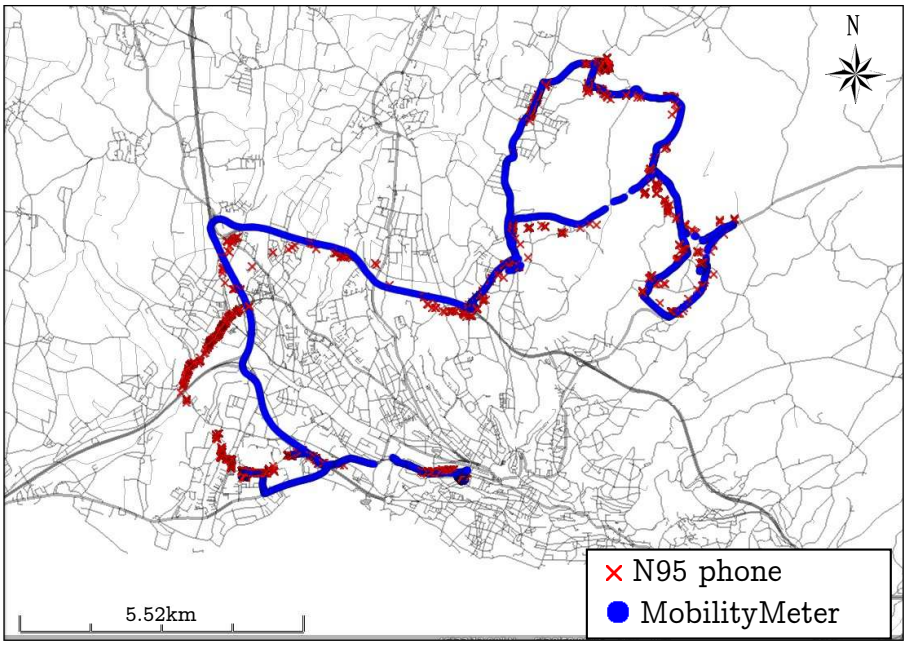
cannot be recorded at a high rate. In our experiments, we use a time interval of 10 seconds. Also, the data is not as accurate as those collected from dedicated GPS devices. For instance, in the Nokia N95 model used for our experiments, the GPS antenna is embedded under the keyboard, which is generally covered by the screen when the phone is not being actively used. Furthermore, most people carry the cell phone in their pocket or handbag. This weakens the GPS signal.

We conducted an experiment where a N95 cell phone and a dedicated GPS device (a MobilityMeter, of the type used by Flamm et al. [2007]) were both carried by the same person during a trip. The two tracks are reported in Figure 1, where the blue circles (appearing darker on a black-and-white copy) represent the tracks provided by the MobilityMeter, and the red x's (appearing lighter on a b&w copy) represent the tracks provided by the N95 smartphone. The significant difference in precision and density appears clearly on these pictures.

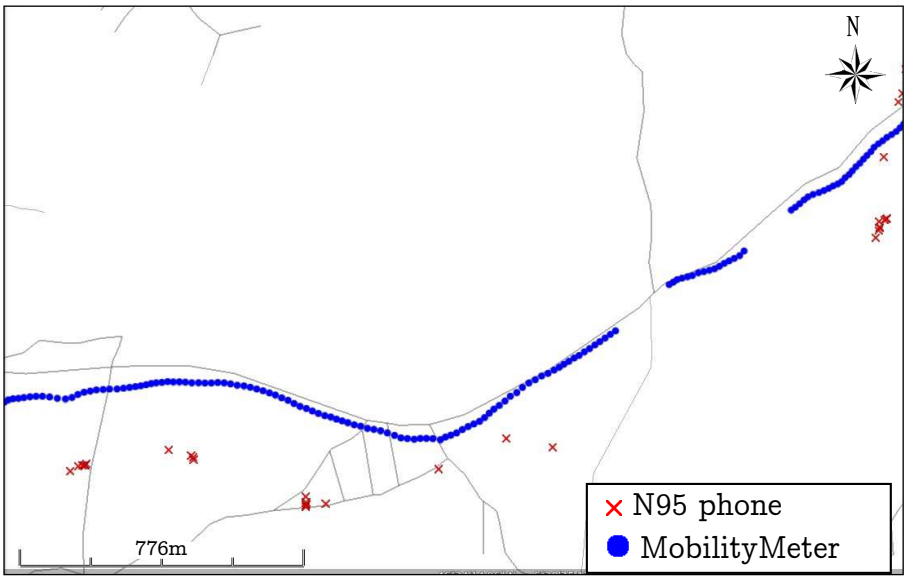
In order to be used as observations for route choice models, the discrete sequences of GPS data need to be linked to the transportation network to generate meaningful paths comprised of connecting arcs.

Map matching (MM) algorithms are generally used to infer from GPS data the corresponding elements in the transportation network, including locations, roads and paths. The main research stream of developing MM is motivated by navigation systems. Consequently, those algorithms aim at providing on-line identification of a real road/arc. A comprehensive review of 35 MM algorithms for navigation applications since 1989 is presented by Quddus et al. [2007]. Many of those algorithms rely not only on GPS data but also on Dead Reckoning (DR) sensors equipping cars or other sensors that smartphones don't embed [e.g., Ochieng et al., 2003, Kim and Kim, 2001]. Because MM algorithms are designed to deterministically detect the correct road for each GPS point, they don't guarantee that detected roads are connected to form a meaningful path, even if some MM algorithms [e.g., Greenfeld, 2002, Ochieng et al., 2003] do consider connectivity and contiguity of the arcs. However, in travel behavior studies, especially in route choice modeling, researchers are not interested in associating every single GPS point to a road. Instead, modelers are interested in the actual path for the whole trip. Incorrectly matched paths may introduce biases in the estimates of the model parameters.

The adaptation of multiple hypotheses technique [Pyo et al., 2001] in



(a) in a region



(b) zoom in

Figure 1: GPS traces from N95 and a GPS device

MM enables modelers to generate a connected path from a GPS trace representing geographical locations during a trip. Developed algorithms [e.g., Marchal et al., 2005, Schuessler and Axhausen, 2009a] maintain at each GPS point a set of path candidates. For each candidate, a score is calculated based on dissimilarity between GPS points and arcs based on distance, speed and/or heading difference, though heading was found to be unreliable for this application [Schuessler and Axhausen, 2009a]. The work by Schuessler and Axhausen [2009a] focuses on the computational efficiency of the MM method, and shows excellent results along that line, with dense and accurate GPS data. However, from experiments that we have conducted (see Section 5), it appears that the method is not suitable for smartphone data, where the focus should be in managing the inaccuracy and low density of the data.

Bierlaire and Frejinger [2008] have introduced an estimation procedure for route choice models that accepts a probabilistic representation of the observed paths, accounting for errors in measurement. An observation does not need to be a unique path, but can be represented by a set of potential paths, along with a probability for each path that it is indeed the actually used path. The scores calculated in MM algorithms, while often heuristically effective, in general lack the theoretical foundation necessary to serve as the probabilities that the corresponding paths are the true path. The simplicity of the score calculation can not ensure its correctness if there are outlier observations. Moreover, in such a post-processing algorithm (as opposed to real time algorithm for navigation tools), "inaccurate" data is eliminated in the process of data filtering [Schuessler and Axhausen, 2009b], with the risk that some useful information is also excluded.

An integrated particle filter modeling framework for detecting transportation modes and traveling roads is proposed by Liao et al. [2007]. In their approach, a *state* combines various mobility patterns, including the transportation mode and the current road. A Rao-Blackwellized particle filter is used as the framework, while the probability of the traveler switching from one mode to another depends on his proximity to available transportation facilities. A Kalman filter is used to model the dynamic process of traveling on the network and retrieving the GPS fix. In order to fit in the Kalman filter framework, a great deal of simplification is required.

In this paper, a method for generating probabilistic path observations from GPS data is proposed. It is capable of dealing with the sparsity and

inaccuracy of the smartphone GPS data, as well as the inaccuracy of the representation of the underlying transportation network. The likelihood that the data has been generated along a given path is calculated.

The next section introduces the GPS data recorded from the smartphones, and the context where the data was recorded. Section 3 derives the model for measuring the likelihood that a GPS trace is recorded while traveling on a path. This model requires a traffic simulator for the underlying transportation network. Although stand-alone traffic simulators can be used, a simple traffic model using only information available from the GPS records are presented in Section 4. Section 5 illustrates the likelihood results calculated for a real trip and four proposed paths, and we also illustrate the MM result for the same data. Potentially true paths need to be generated before their likelihoods can be calculated. However theoretical and numerical analysis reveal that traditional MM algorithms are not suitable for the smartphone GPS data. Therefore a new path generation algorithm, accounting for the sparsity of the smartphone GPS data, is proposed in Section 6. In Section 7, the driving route choice behavior of a smartphone user is modeled from path observations, which are generated from real GPS data using the proposed methods. Finally, some conclusions are included in section 8.

2 Context and data

Let $G = (N, A)$ denote a transportation network, where N is the set of all nodes and A the set of all arcs. The horizontal position of each node $n \in N$ is represented by $x_n = \{\text{lat}, \text{lon}\}$, which is a pair of coordinates consisting of latitude and longitude. The shape of the physical route of arc a is described by an application

$$\mathcal{L}_a : [0, 1] \to \mathbb{R}^2. \tag{1}$$

For a point on the arc, its position x is generated from a unique number ℓ between 0 and 1 such that $x = \mathcal{L}_a(\ell)$. In particular, $\mathcal{L}_a(0)$ is the coordinates of the up-node, and $\mathcal{L}_a(1)$ is the coordinates of the down-node of arc a. For example, if the arc is a straight line from node u to node d, then

$$\mathcal{L}_a(\ell) = (1 - \ell)\, x_u + \ell x_d. \tag{2}$$

The performance of the network is characterized by a model

$$x = S(x^-, t^-, t, p) \tag{3}$$

predicting the position x at time t of an individual being in position x^- at time t^-, and following path p. It is a random variable with probability distribution function

$$f_x(x|x^-, t^-, t, p). \tag{4}$$

Typically, this model is obtained from a calibrated traffic simulator. However, for practical purposes, analytical models can also be used (see Section 4).

Location data is recorded by devices which are carried by travelers when they are traveling in the transportation network. The device makes location measurements combining various sensors such as GPS readings, GSM cell tower information, WLAN base stations, etc. We denote one measurement by

$$\widehat{g} = \left(\widehat{t}, \widehat{x}, \widehat{\sigma}^x, \widehat{v}, \widehat{\sigma}^v, \widehat{h}\right),$$

which is a tuple containing:

- $\widehat{t}$, a time stamp ;
- $\widehat{x} = (\widehat{x}_{lat}, \widehat{x}_{lon})$, a pair of coordinates;
- $\widehat{\sigma}^x$, the standard deviation of the horizontal error in the location measurement;
- $\widehat{v}$, a speed measurement (km/h) and,
- $\widehat{\sigma}^v$, the standard deviation of the error in that measurement;
- $\widehat{h}$, a heading measurement, that is the angle to the north direction, from 0 to 359, clockwise.

We assume that the data has been preprocessed so that we have access to a sequence of measurements $(\widehat{g}_1, \ldots, \widehat{g}_T)$ corresponding to a given trip.

3 Matching paths with GPS data

In this section, we focus on the computation of the likelihood of a set of GPS data. More precisely, given a sequence of measurements $(\widehat{g}_1, \ldots, \widehat{g}_T)$ and a path p, we compute the probability that the true path that was followed by the data generating device was actually p.

We first describe some score functions proposed in the literature. In Section 3.2, we introduce a new modeling framework to derive the probability.

3.1 State of the art

In the Multiple Hypotheses Technique (MHT) based MM algorithm proposed by Marchal et al. [2005], a score is calculated for each candidate p based on dissimilarities between GPS points $(\widehat{g}_1, \ldots, \widehat{g}_T)$ and arcs on path p based on distance:

$$SC_p = \sum_{a \in p} \sum_{k=1}^{T} d(\widehat{x}_k, a)\, \delta_{ak}, \tag{5}$$

where SC_p is the score of path p, $\delta_{ak} = 1$ if $\widehat{g}_k$ is matched by arc a (see below), 0 otherwise, and $d(\widehat{x}, a)$ represents the perpendicular distance between the GPS point $\widehat{g}$ and the matched arc a. The perpendicular distance is defined as the euclidean distance between $\widehat{x}$ and its projection $\widehat{x}'$ on the line supporting arc a, if $\widehat{x}'$ lies on arc a. Otherwise, it is defined as the euclidean distance between $\widehat{x}$ and the start node, or the end node, depending on which one is the smallest. As a result, the path with the lowest score is selected as the "true" one. Schuessler and Axhausen [2009a] extend this method by discounting score if the observed speed of a GPS point exceeds the free-flow speed on the matched arc, and propose the following specification:

$$SC_p = \sum_{a \in p} \sum_{k=1}^{T} \left(d(\widehat{x}_k, a)\, \delta_{ak} + (\widehat{v}_k - v_{ff}(a))^2 \gamma_{ak} \right), \tag{6}$$

where $v_{ff}(a)$ is the free flow speed on arc a; $\gamma_{ak} = 1$ if the observed speed is larger than the free flow speed, that is, $\widehat{v}_k > v_{ff}(a)$, and 0 otherwise. Although the speed penalty term is reasonable, Schuessler (private communication) did not observe a major influence of that part of the score

on the results with her data. Also, the free flow speed data is difficult to obtain in practice.

In the integrated particle filter modeling framework proposed by Liao et al. [2007] for detecting transportation modes and traveling arcs from GPS points, the MM is modeled by a Kalman filter. The state of the system is defined as a combination of mobility features, including transportation mode, location, and speed. It involves a model for system dynamics (a structural equation), predicting the state at time k knowing the state at time $k-1$, as well as a sensor model (a measurement equation) providing the likelihood of a data point. Both models are assumed to be simple Gaussian models with given covariance structures.

In our our approach, we also propose a framework based on measurement and structural equations, but we derive these equations differently. The structural equation is provided by the traffic model (3) and the associated distribution (4). They are discussed more intensively in Section 4. The measurement equations are derived in the next subsection.

3.2 Measurement equations

We now derive the probability that a given path p generates the data $(\widehat{g}_1, \ldots, \widehat{g}_T)$. For the sake of simplification, we focus on the measurement equation for the locations $(\widehat{x}_1, \ldots, \widehat{x}_T)$, that is

$$\Pr(\widehat{x}_1, \ldots, \widehat{x}_T | p), \tag{7}$$

which is decomposed recursively:

$$\Pr(\widehat{x}_1, \ldots, \widehat{x}_T | p) = \Pr(\widehat{x}_T | \widehat{x}_1, \ldots, \widehat{x}_{T-1}, p) \Pr(\widehat{x}_1, \ldots, \widehat{x}_{T-1} | p). \tag{8}$$

The recursion starts with the model $\Pr(\widehat{x}_1|p)$:

$$\Pr(\widehat{x}_1|p) = \int_{x_1 \in p} \Pr(\widehat{x}_1|x_1, p) \Pr(x_1|p) dx_1, \tag{9}$$

where the integral spans all locations x_1 on path p. For the first point, we do not have any prior on the location, and therefore, $\Pr(x_1|p)$ is a constant equal to the inverse of the length L_p of p. The model $\Pr(\widehat{x}_1|x_1, p) = \Pr(\widehat{x}_1|x_1)$ describes the measurement error of the smartphone device. For instance, we may assume that it follows a Rayleigh distribution, which is derived from the assumption that the latitudinal and longitudinal errors are i.i.d. normal

with variance σ^2. As σ^2 is unknown, we use $\widehat{\sigma}^2 = \sigma^2_{\text{network}} + (\widehat{\sigma}^x_1)^2$ as an estimate, where $\sigma^2_{\text{network}}$ captures the difference between the coded network and the actual roads and paths, and $(\widehat{\sigma}^x_1)^2$ captures the measurement error of the GPS device. Therefore,

$$\Pr(\widehat{x}_1|x_1) = \exp\left(-\frac{\|\widehat{x}_1 - x_1\|_2^2}{2\widehat{\sigma}^2}\right). \tag{10}$$

Combining (9) and (10), we obtain

$$\Pr(\widehat{x}_1|p) = \frac{1}{L_p}\int_{x_1} \exp\left(-\frac{\|\widehat{x}_1 - x_1\|_2^2}{2\widehat{\sigma}^2}\right) dx_1. \tag{11}$$

For long paths, this integral may be cumbersome to compute. In this case, we propose to simplify its computation using the concept of Domain of Data Relevance (DDR) introduced by Bierlaire and Frejinger [2008], as described in Section 3.3.

The next step of the recursion derives

$$\Pr(\widehat{x}_1, \widehat{x}_2|p) = \Pr(\widehat{x}_2|\widehat{x}_1, p)\Pr(\widehat{x}_1|p), \tag{12}$$

where $\Pr(\widehat{x}_1|p)$ is defined by (11). We write

$$\Pr(\widehat{x}_2|\widehat{x}_1, p) = \int_{x_2 \in p} \Pr(\widehat{x}_2|x_2, \widehat{x}_1, p)\Pr(x_2|\widehat{x}_1, p)dx_2. \tag{13}$$

The first term in (13), $\Pr(\widehat{x}_2|x_2, \widehat{x}_1, p) = \Pr(\widehat{x}_2|x_2)$, is again modeling the measurement error of the device, and can also be defined by (10), combined with the same simplifications as described above. The second term predicts the position at time $\widehat{t}_2$ of the traveler. It is written as

$$\Pr(x_2|\widehat{x}_1, p) = \int_{x_1 \in p} \Pr(x_2|x_1, \widehat{x}_1, p)\Pr(x_1|\widehat{x}_1, p)dx_1. \tag{14}$$

The first term in (14) models the movement of the traveler, which is captured by (3), that is

$$\Pr(x_2|x_1, \widehat{x}_1, p) = f_x(x_2|x_1, \widehat{t}_1, \widehat{t}_2, p),$$

where f_x is the density function (4) of the traffic model. The second term can be derived from Bayes rule:

$$\Pr(x_1|\widehat{x}_1, p) = \frac{\Pr(\widehat{x}_1|x_1, p)\Pr(x_1|p)}{\int_{x_1}\Pr(\widehat{x}_1|x_1, p)\Pr(x_1|p)dx_1}.$$

As $\Pr(x_1|p) = 1/L_p$ is constant for a given p, we have

$$\Pr(x_1|\widehat{x}_1, p) = \frac{\Pr(\widehat{x}_1|x_1, p)}{\int_{x_1} \Pr(\widehat{x}_1|x_1, p)dx_1} \tag{15}$$

which is a normalized version of (10). This completes the definition of (12).

The recursion in (8) requires that, at iteration k, the probability

$$\Pr(\widehat{x}_k|\widehat{x}_1, \ldots, \widehat{x}_{k-1}, p)$$

is calculated. It can be generalized from (13) and (14) that

$$\Pr(\widehat{x}_k|\widehat{x}_1, \ldots, \widehat{x}_{k-1}, p) = \int_{x_k} \Pr(\widehat{x}_k|x_k, \widehat{x}_1, \ldots, \widehat{x}_{k-1}, p) \int_{x_{k-1}} \Pr(x_k|x_{k-1}, p) \Pr(x_{k-1}|\widehat{x}_1, \ldots, \widehat{x}_{k-1}, p) d_{x_{k-1}} dx_k, \tag{16}$$

where $\Pr(\widehat{x}_k|x_k, \widehat{x}_1, \ldots, \widehat{x}_{k-1}, p) = \Pr(\widehat{x}_k|x_k)$ is given by (10), and $\Pr(x_k|x_{k-1}, p)$ is the traffic model $f_x(x_k|x_{k-1}, \widehat{t}_{k-1}, \widehat{t}_k, p)$. The last part of (16), $\Pr(x_{k-1}|\widehat{x}_1, \ldots, \widehat{x}_{k-1}, p)$, is the posterior pdf of the true location x_{k-1} given observed GPS trace $\widehat{x}_1, \ldots, \widehat{x}_{k-1}$ and path p. This distribution is not tractable, and we must simplify it, and replace it by

$$\Pr(x_{k-1}|\widehat{x}_1, \ldots, \widehat{x}_{k-1}, p) \approx \Pr(x_{k-1}|\widehat{x}_{k-1}, p). \tag{17}$$

Therefore, we can use the same derivation that leads to (15) to obtain

$$\Pr(x_{k-1}|\widehat{x}_{k-1}, p) = \frac{\Pr(\widehat{x}_{k-1}|x_{k-1}, p)}{\int_x \Pr(\widehat{x}_{k-1}|x, p)dx}. \tag{18}$$

The derivation above involves many integrals over the full path. Although these integrals have low dimension, they can be cumbersome to compute, especially when the path p is long. In the next section, we describe how to decompose the integrals, and to use the concept of Domain of Data Relevance (DDR) introduced by Bierlaire and Frejinger [2008] to simplify the computation.

3.3 Computing integrals

The measurement equations involve various integrals along a path p of the form

$$I = \int_{x \in p} f(x)dx, \tag{19}$$

that are complicated to compute in real applications. We describe here how to exploit the topology of the network to compute these integrals.

First, we decompose the path into arcs to obtain

$$I = \sum_{a \in p} \int_{x \in a} f(x) dx. \tag{20}$$

For each arc, we use the shape model (1) to obtain a unidimensional integral

$$\int_{x \in a} f(x) dx = \int_{\ell=0}^{1} f(\mathcal{L}_a(\ell)) |\partial \mathcal{L}| d\ell, \tag{21}$$

where

$$|\partial \mathcal{L}| = \sqrt{\left(\frac{d(\mathcal{L}_a(\ell))_{lat}}{d\ell}\right)^2 + \left(\frac{d(\mathcal{L}_a(\ell))_{lon}}{d\ell}\right)^2}. \tag{22}$$

For example, if the linear model (2) is used, we have

$$|\partial \mathcal{L}| = \|x_u - x_d\|_2. \tag{23}$$

Second, we truncate the domain of the integrals to save computation time where negligible quantities are involved. For a given GPS observation $\widehat{x}$, Bierlaire and Frejinger [2008] define the DDR as the physical area where the piece of data is relevant. In our context, a point x is considered to be in the DDR of $\widehat{x}$ if the probability $Pr(\widehat{x}|x)$ is above a given threshold θ, and the heading difference between the GPS point and the arc is less than 60 degrees if $\widehat{v} > 8km/h$.[1] In our implementation, we have used a value $\theta = 0.65$. It corresponds roughly to points in a diameter of 100m when the σ parameter of the GPS device is 100m, and the σ for the network coding is assumed to be 30m. Indeed,

$$\exp\left(-\frac{\|\widehat{x} - x\|_2^2}{2\widehat{\sigma}^2}\right) \geq \theta$$

is equivalent to

$$\|\widehat{x} - x\|_2 \leq \sqrt{-2\left(\widehat{\sigma}\right)^2 \ln \theta},$$

and the upper bound 96.9 is obtained with $\theta = 0.65$ and $\widehat{\sigma} = 104.4 = \sqrt{100^2 + 30^2}$. This is illustrated in Figure 2, where the parts of arcs AB

[1] At low speeds, heading measurements from the GPS are generally not reliable.

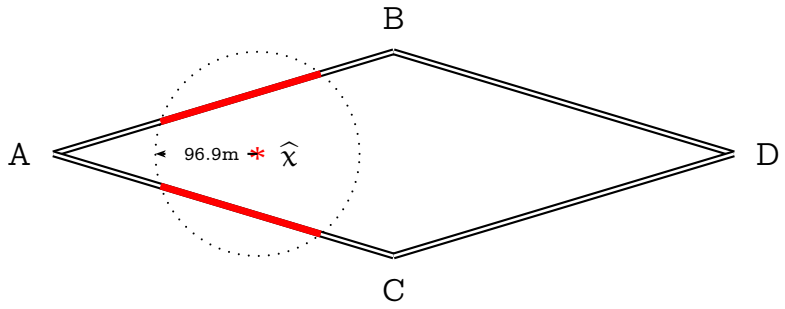


Figure 2: Domain of Data Relevance

and AC represented by a solid red line are inside the DDR of the data point $\widehat{x}$.

Clearly, the value of the parameters should be adjusted to account for the features of the relevant application, and the quality of the associated data. Also, the complexity of the computation of the integrals increases with the size of the DDR. A large DDR means more computation. On the other hand, too small a DDR may artificially produce a zero probability for the measurement equation, which is undesirable. As discussed by Bierlaire and Frejinger [2008], the specification of the DDR should correspond to a good trade-off between accuracy and computational burden.

4 Traffic model

In our framework, the traffic model is designed to predict the position of the GPS device over time. More precisely, it predicts the position x of the device at time t if the position at time t^- is x^-, and the device is traveling along path p.

This is the typical role of dynamic traffic simulators (such as AIMSUN Barceló and Casas [2005], MITSIM Yang and Koutsopoulos [1996], Dyna-MIT Ben-Akiva et al. [2001], Dynasmart Mahmassani [2001], among many). However, it is not always practical to use a calibrated traffic simulator in a MM context. Therefore, we suggest to use simple analytical models such as the one described below.

In order to derive the traffic model (4), we define the operator that computes the distance between two points x and y lying on path p, and denote it by

$$d_p(x, y). \tag{24}$$

parameter	estimate	standard error
w	0.528	0.0362
λ	0.041	0.0032
μ	3.843	0.0206
τ	0.250	0.0200
Parameters estimated by R.		

Table 1: Parameters estimates for the speed distribution

This operator is easily implemented using the same decomposition of paths into arcs described in Section 3.3. We write the traffic model in terms of speed instead of position, considering the random variable

$$v = \frac{d_p(x^-, x)}{t - t^-} \tag{25}$$

with pdf

$$f_v\left(\frac{d_p(x^-, x)}{t - t^-}\right). \tag{26}$$

In our experiments, the traveling speed of the device is recorded every 10 seconds, therefore its distribution can be derived from the observed speed data. For the distribution of speed, we assume a mixture of a negative exponential distribution and a log normal distribution. The first is designed to capture the instances where the vehicles are stopped at intersections, or traveling at low speed before or after that stop. The second is designed to capture vehicles moving at regular speed. The distribution is

$$f_v(v) = w\lambda \exp^{-\lambda v} + (1 - w)\frac{1}{v\sqrt{\pi\tau^2}} \exp^{-\frac{(\ln v - \mu)^2}{2\tau^2}}, \tag{27}$$

where w (the weighting), λ (the scale parameter of the negative exponential distribution), μ (the location parameter of the log normal distribution), and τ (the scale parameter of the log normal distribution) are parameters to be estimated. The data for the estimation consists of 658 speed records observed from a user while he was traveling. Figure 3 shows the normalized histogram of the recorded speed data and the estimated speed distribution. Table 1 reports the parameters estimated by maximum likelihood.

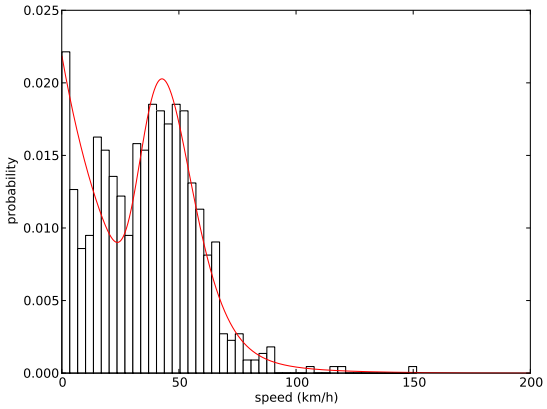


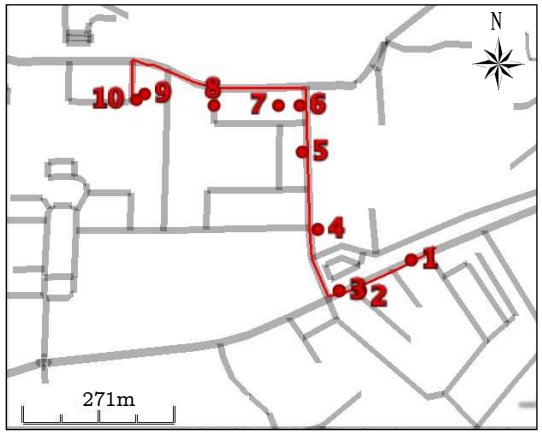
Figure 3: The speed distribution

5 Illustration

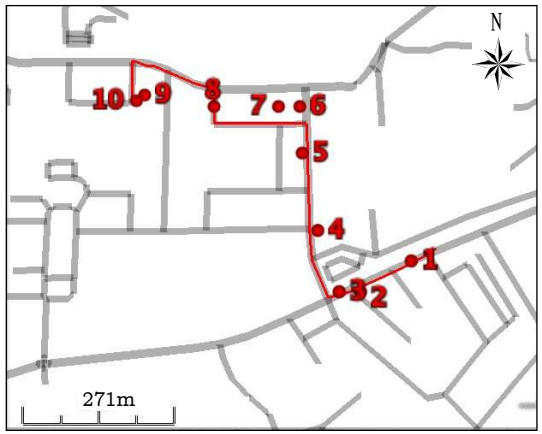
This section illustrates the likelihood result calculated for a trace of real GPS data (shown as red points in Figure 4, appearing in gray on a black and white copy) using the methodologies described in Section 3 and Section 4. This GPS trace was recorded from a Nokia 95 smartphone recording data points at 10 second intervals. The coordinates are recorded in WGS84 format. Detailed information about the GPS readings has been introduced in Section 2.

This particular GPS trace was chosen to be analyzed because it was recorded while traveling by car in a dense transportation network. The actual path is shown in Figure 4(a) as the solid red line, and is known with certainty as the traveler was one of the authors. The ambiguity of the coordinates readings and the density of the transportation network makes the actual path difficult to be recognized from the data alone. The transportation network data used is provided by openstreetmap (`www.openstreetmap.org`), which is an open source map data service.

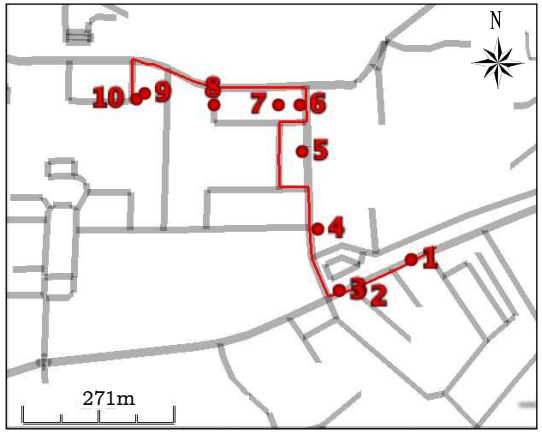
It can be observed from Figure 4(a) that some of the GPS points (e.g. 7 and 8) deviate more than 30 meters from the actual path. Consequently, another path shown in Figure 4(b) also seems intuitively reasonable enough to be the actual path if we only compare the geographical dissimilarities. The natural logarithm of the measurement likelihood (7), termed the mea-



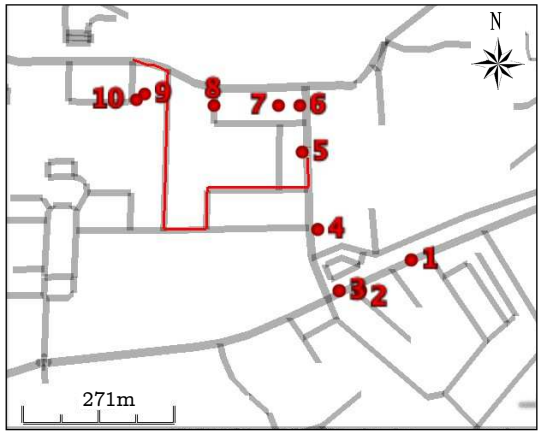
(a) The actual path (−11.3)



(b) Alternative candidate 1 (−12.9)



(c) Alternative candidate 2 (−13.2)



(d) Map matching path (0)

Figure 4: A real GPS trace

surement log likelihood[2],

$$\ln \Pr(\widehat{x}_1, \ldots, \widehat{x}_T | p) \tag{28}$$

for the actual path is -11.3, while it is only -12.9 for the other path. Another path candidate shown in Figure 4(c) is intuitively less possible to be the actual path, and the log likelihood for this path is lower still, -13.2.

We also apply the MM algorithm developed by Schuessler and Axhausen [2009a], and generate a deterministic MM path shown in Figure 4(d). This result looks strange due to the incapability of the algorithm to deal with sparse data, as described in Section 6 in detail. And the measurement likelihood value (7) for this path is 0, because the path doesn't pass through DDRs of some GPS points (e.g. 1).

6 Path generation algorithm

For a set of GPS data, the method presented in Section 3 assigns a likelihood to a given path p. We focus now on the path generation process itself.

State of the art algorithms are designed for dense data, where it can be safely assumed that nearly every arc on a path generates at least one GPS point. For instance, Marchal et al. [2005], Schuessler and Axhausen [2009a] generate path candidates by considering each GPS point one by one in the chronological order. At each iteration k, they generate a set P_k of path candidates assumed to match the GPS points up to k. They generate new candidates by topologically extending the paths in P_{k-1}, and select a fixed number of them to belong to P_k according to the score function described in Section 3.1.

It can clearly be observed from Figure 1 that the dedicated GPS device data is consistent with the "high density" hypothesis, while the smartphone data is not. Also, the example shown in Figure 4(d) shows that the MM algorithm is not appropriate for smartphone GPS data.

In order to address this problem, we propose a path generation algorithm designed for sparse data. First, we identify GPS points that have a

[2]If we further expand (8), the measurement likelihood (7) becomes $\Pr(\widehat{x}_1, \ldots, \widehat{x}_T | p) = \Pr(\widehat{x}_1 | p) \prod_{k=2}^{T} \Pr(\widehat{x}_k | \widehat{x}_1, \ldots, \widehat{x}_{k-1}, p)$, which is the multiplication of many probability values that are smaller than 1. Consequently, the measurement likelihood (7) is close to zero. Throughout this paper we present the logarithm of it, if it is not zero.

speed lower than 8km/h as "stationary". When the device is more or less stationary, while it may generate data that is relevant for comparing path likelihood, it is not generating information that is useful in path generation. Two exceptions are the first and the last GPS points; even if their speed values are low, they reveal information about the origin and the destination.

The key idea is to associate a DDR D_k and a set of arcs L_k with each relevant GPS point $\widehat{g}_k$, and to generate the path set P_k from L_k. At each iteration k, the algorithm performs three steps.

1. Bounded shortest path trees are generated from the end nodes of each path in P_{k-1}. The bound is derived from an assumption about the maximum possible speed and the time interval between t_{k-1} and t_k. The leaf nodes of the bounded shortest path tree are the first nodes detected by the Dijkstra algorithm that violate the bound. In our experiments, the bound is defined by $1.5(t_k - t_{k-1})\widehat{v}_{max}$, where $\widehat{v}_{max}$ is the maximum speed value among the observed speeds $\widehat{v}_{k-1}$ and $\widehat{v}_k$, and the speed calculated by $\|\widehat{x}_k - \widehat{x}_{k-1}\|_2/(t_k - t_{k-1})$, and the factor 1.5 is a safety margin to minimize the risk of missing a relevant observation.

2. The DDR D_k associated with the data point $\widehat{g}_k$ is constructed using the conditions described in Section 3.3. An arc belongs to the set L_k if it belongs to one of the bounded shortest path tree generated during the first step, and it intersects with D_k. Clearly, at the first iteration, only the latter condition applies.

3. Each path in P_{k-1} is now extended by connecting it with all arcs in L_k, to generate a candidate set P'_k. For a given path p in P_{k-1} and a given arc a in L_k, the shortest path between the end node of p and the up node of a is appended to p (note that it is simply extracted from the bounded shortest path trees generated above). The extension takes place only if the first arc of the shortest path belongs to D_k or is not the reverse arc of the connecting arc. This is designed to exclude unreasonable U-turns. The resulting path is included in P'_k. If the size of the set P'_k becomes large (say, larger than 20), not all paths are kept into P_k. The following selection procedure is applied.

   (a) The 2 shortest paths in P'_k are selected.

(b) Paths are randomly selected from P'_k according to the likelihood (7). In practice, the likelihood is normalized to obtain scores summing up to one in P'_k before the random selection. Path candidates are drawn and included in P_k using simulation until the cumulative normalized likelihood exceeds a predefined number (e.g. 0.8). Note that (7) is computed with all GPS data, including the stationary points.

(c) For each arc a in L_k, we define P'_{ak} as the set of paths in P'_k containing a. We then apply the same simulation procedure on P'_{ak}. This is meant to guarantee that each arc associated with the latest GPS point has a path associated to it.

The result of this procedure is a set of paths P, and in the following, we assume that the actual path that the smartphone user travels on belongs to this set. Consequently, the set of OD pairs S associated with P is the set of all potentially true OD pairs.

We illustrate the effect of the proposed path generation algorithm by applying it to a real trip with 53 GPS points (see Figure 5). The trip direction is from west to east. The algorithm generates 8 candidate paths, associated with 3 OD pairs. Table 2 illustrates details of the path candidates, including the origins and the destinations' identifiers (id) coded in the network data, the measurement log likelihood (28), the length in kilometers, and the number of traffic signals. It also contains the approximate travel time of the trip, which is the time difference between the first and the last GPS points, 533 seconds. These paths show substantial overlap. For example, the result reveals that, at the end of the trip (zoomed in the Figure), the traveler either traveled through the main road (path 1-3), to which the GPS points are close, or made a detour to residential roads (path 4-8), which are parallel and west of the main road. The result reveals that the GPS measurement is more likely to be on the paths that go via the main roads, because as shown in the table, the measurement log likelihood (28) for path 1-3 is much higher than that for path 4-8. This result seems intuitively reasonable, although we don't have access to the actual path to validate it.

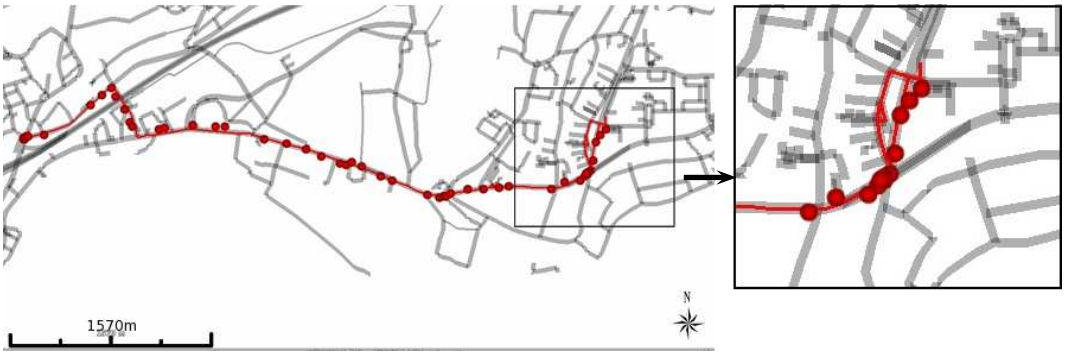


Figure 5: Illustration of the path generation algorithm

path id	origin id	destination id	log likelihood (28)	length (km)	traffic signals
1	252978965	253301632	-59.62	5.57	5
2	252978965	388419949	-59.45	5.52	5
3	252978965	253301629	-59.92	5.49	5
4	252978965	253301629	-62.30	5.60	5
5	252978965	253301629	-62.55	5.63	5
6	252978965	253301629	-62.63	5.64	5
7	252978965	253301629	-62.29	5.60	5
8	252978965	253301629	-62.63	5.64	5
Number of GPS points: 53, approximate travel time: 553 seconds					

Table 2: Details of a trip.

7 Modeling route choice behavior from real data

The motivation to develop the probabilistic path generation algorithm is the estimation of route choice models. In this section, we illustrate the estimation of the parameters of a route choice model for a given smartphone user from data recorded from his phone while driving. We use the network-free data modeling and the Path Size Logit (PSL) as the route choice model [Bierlaire and Frejinger, 2008]. For a GPS trace $\{\widehat{x}_1, \ldots, \widehat{x}_T\}$ that represents a trip, P and S are the generated potential true paths and OD pairs. The likelihood function for this GPS trace is given by

$$\Pr(\widehat{x}_1, \ldots, \widehat{x}_T|S) = \sum_{s \in S} \Pr(s|S) \sum_{p \in P^s} \Pr(\widehat{x}_1, \ldots, \widehat{x}_T|p) \Pr(p|C(s); \beta), \quad (29)$$

where

- S is the set of relevant OD pairs,
- $\Pr(s|S)$ is the probability that the actual OD pair is s. In this study, it is defined as $\Pr(s|S) = 1/|S|$ if $s \in S$, and 0 otherwise;
- $P^s \subseteq P$ is the set of generated path candidates corresponding to OD pair $s \in S$;
- $\Pr(\widehat{x}_1, \ldots, \widehat{x}_T|p)$ is the GPS measurement likelihood (7) calculated from the proposed method;
- $\Pr(p|C(s); \beta)$ is the route choice model, where $C(s)$ is the choice set for OD pair s, and β are the parameters to be estimated. In this study, a PSL specification is used.

In the following subsections, we focus on the specification and the estimation of the choice model $\Pr(p|C(s); \beta)$, where $p \in P^s$ is a probabilistically chosen path with OD pair s.

7.1 Choice set generation: importance sampling

Choice set generation is an important procedure in route choice modeling. In this study, we employ the stochastic choice set generation algorithm proposed by Frejinger et al. [2009]. This method assumes that the relevant choice set $C(s)$ is the set of all possible paths in the network connecting

OD pair s. In order to develop a tractable choice set for use in estimating the parameters of the choice model, path alternatives are sampled using a biased random walk algorithm, with arc weights at each node set by the ratio of the length of the shortest path to the destination using any arc and using the target arc. The sampling bias is subsequently corrected in the choice model.

The random walk procedure as presented in Frejinger et al. [2009] does not allow passing the destination and subsequently returning to it. However, GPS observations show this behavior is not uncommon, as it can represent normal parking search behavior. In order to incorporate the sampling bias correction in the choice model, the positive conditioning property requires that it is at least possible to sample the observed path choice, so we modify the random walk algorithm to allow for this.

Let h be the number of times the random walk algorithm has visited the destination node. At each such visit, the walk terminates with probability $P(h)$, where $P(h)$ is increasing with h, and the walk continues with probability $1 - P(h)$. In our experiments, we used

$$P(h) = 1 - 0.5^h. \tag{30}$$

If the walk proceeds, it does so using the original procedure, although we modify the arc weights for selecting an arc departing the destination, to reflect the fact that the walk has to leave the destination node. Otherwise, the shortest path has zero length, resulting in undefined arc selection probabilities. We correct this by simply imposing a condition that a shortest path must contain at least one arc, thus forcing a strictly positive result.

With this modification, the probability $q(p)$ for sampling a path p is now

$$q(p) = P(h_p) \prod_{i=1}^{h_p - 1} (1 - P(i)) \prod_{a \in \Gamma_p} q(a|\mathcal{E}_a),$$

that is, the probability that the algorithm has been continued $h_p - 1$ times and stopped once. Γ_p is the (ordered) sequence of arcs in path p, $\mathcal{E}_a$ is the list of arcs with the same up-node as a[3], $q(a|\mathcal{E}_a)$ is the probability for

[3]Note that we use a slightly different notation than Frejinger et al. [2009] to simplify the presentation.

	Min	Average	Max
Number of GPS points per trip	16	36	58
Approximate travel time per trip [second]	179	397	795
Length of the generated paths [km]	1.93	3.98	6.42
Number of traffic signals of the generated paths	0	2.84	5.0

Table 3: Statistics of the recorded 19 trips.

selecting arc a among all arcs in $\mathcal{E}_a$, h_p is the number of times that the destination node is in path p.

To estimate the models presented in the next section, choice set samples were created by generating 50 random walks for comparison against each possible true path, using the method described above with length representing generalized cost and Kumaraswamy parameters b_1=30 and b_2=1, plus the observed paths as calculated by the previously described algorithms (see Ben-Akiva and Lerman [1985] for a discussion of model estimation in general).

7.2 Model estimation

We estimate the parameters of a simple model in order to illustrate the procedure. We use 19 real trips recorded from a single user's smartphone. Table 3 presents some statistics about the trips, as well as about the paths generated by the procedure described in Section 6. For each trip, the length and the number of traffic signals are weighted by the normalized measurement likelihood:

$$\Pr(p|\widehat{g}_1,\ldots,\widehat{g}_T) = \frac{\Pr(\widehat{g}_1,\ldots,\widehat{g}_T)|p}{\sum_{p'\in P}\Pr(\widehat{g}_1,\ldots,\widehat{g}_T|p')}. \tag{31}$$

The deterministic term of the utility function of path p in the PSL model is specified as

$$V_p = \beta_{EPS}\ln EPS_p + \beta_\ell L_p + \beta_{sg}\mathrm{NbSignals} + \mathrm{Corr}_p, \tag{32}$$

where $\mathrm{NbSignals}$ is the number of traffic signals along the path; EPS_p is the Extended Path Size (EPS), which accounts for the path overlapping and corrects for the sampling; Corr_p is the choice set sampling correction term. We refer to Frejinger et al. [2009] for more details about EPS and sampling correction.

Coefficient	Value	Rob. Std. Error	Rob. t-test	p value
β_{EPS}	0.242	0.138	4.98	0.00
β_{ℓ}	-33.7	16.4	-5.28	0.00
β_{sg}	-2.74	3.67	-2.39	0.02
Number of observations: 19				
Null log likelihood: -776.1				
Final log likelihood: -708.9				
Adjusted rho-square: 0.083				
Model estimated by BIOGEME [Bierlaire, 2003]				

Table 4: Estimation result.

Table 4 reports the coefficient estimates. All coefficients have their expected signs (positive for the EPS coefficient as is consistent with established route choice theory [Frejinger, 2008], and negative for coefficients on path length and number of traffic signals) and they are all significantly different from zero.

Clearly, the size of the sample is too small to consider this as a final model. However, it illustrates the feasibility of the overall approach on a real data set.

8 Conclusions

In this paper, we propose a methodology to estimate route choice models from GPS data. It builds on the work by Frejinger [2008], Bierlaire and Frejinger [2008], Frejinger et al. [2009]. We introduce a systematic method for matching a set of paths with GPS data. A measurement equation is derived, which calculates the probability that the device would have generated a sequence of GPS tracks while following a given path. It is based on a structural model, which captures the movements of the GPS device. By simulating both the travel dynamics and the recording of the traveler in the transportation network using these two models, the uncertainty derived from the inaccuracy of both the GPS data and the transportation network is taken into account. The application to real data shows that the probability values of the actual path and some other paths are realistic and meaningful. Moreover, the comparison against a state of the art MM algorithm shows that it is particularly suitable for smartphone GPS data, which are typically

sparse and inaccurate.

A path generation algorithm is also proposed that accounts for the sparsity of the data. The methodology has been applied on real smartphone data collected in Switzerland. The estimation of a simple route choice model from real data illustrates that the proposed methodology indeed allows the use of GPS data from smartphones.

Future extensions of this work include investigation into the use of other types of data provided by smartphones, such as the detection of cell towers, WiFi base stations, or other bluetooth devices, as well as physical activity detected by accelerometers, gyroscopes, and magnetometers. Also, the analysis of other travel decisions, such as mode choice, should be considered, similar to the work by Liao et al. [2007]. Moreover, we are interested in motivating and developing a sampling protocol where all generated path candidates are considered simultaneously within the choice model $\Pr(p|C(s);\beta)$. Finally, the efficiency of the algorithm will have to be adapted to deal with large networks.

9 Ackowledgments

We have benefited from discussions with Emma Frejinger and the members of the Nokia Research Center in Lausanne, especially Niko Kiukkonen, whose help in data collection was invaluable. The earlier part of this research was founded by Nokia. The second author is supported by the Swiss National Science Foundation grant 200021/131998 *Route choice models and smart phone data*.

References

J. Barceló and J. Casas. Dynamic network simulation with aimsun. *Simulation Approaches in Transportation Analysis*, pages 57–98, 2005. URL `http://dx.doi.org/10.1007/0-387-24109-4_3`.

M. Ben-Akiva and S. R. Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press Series in Transportation Studies. The MIT Press, Cambridge, MA, 1985.

M. Ben-Akiva, M. Bierlaire, D. Burton, H. Koutsopoulos, and R. Mishalani.

Network state estimation and prediction for real-time traffic management. *Networks and Spatial Economics*, 1(3):293–318, 09 2001. URL http://dx.doi.org/10.1023/A:1012883811652.

M. Bierlaire. Biogeme: a free package for the estimation of discrete choice models. In *3rd Swiss Transportation Research Conference*, Ascona, Switzerland, 2003.

M. Bierlaire and E. Frejinger. Route choice modeling with network-free data. *Transportation Research Part C: Emerging Technologies*, 16 (2):187–198, 2008.

M. Flamm, C. Jemelin, and V. Kaufmann. Combining person based GPS tracking and prompted recall interviews for a comprehensive investigation of travel behaviour adaptation processes during life course transitions. In *7th Swiss Transport Research Conference*, Ascona, Switzerland, 2007.

E. Frejinger. *Route choice analysis: data, models, algorithms and applications*. PhD thesis, EPFL, 2008.

E. Frejinger, M. Bierlaire, and M. Ben-Akiva. Sampling of alternatives for route choice modeling. *Transportation Research Part B: Methodological*, 43(10):984–994, 2009.

J. Greenfeld. Matching GPS observations to locations on a digital map. In *Proceedings of the 81th Annual Meeting of the Transportation Research Board*, Washington, D.C., USA, 2002.

S. Kim and J.-H. Kim. Adaptive fuzzy-network-based c-measure map-matching algorithm for car navigation system. *IEEE transactions on industrial electronics*, 48(2):432–441, 2001.

L. Liao, D. J. Patterson, D. Fox, and H. Kautz. Learning and inferring transportation routines. *Artificial intelligence*, 171(5-6):311–331, 2007.

H. Mahmassani. Dynamic network traffic assignment and simulation methodology for advanced system management applications. *Networks and Spatial Economics*, 1(3):267–292, 09 2001. URL http://dx.doi.org/10.1023/A:1012831808926.

F. Marchal, J. Hackney, and K. Axhausen. Efficient map matching of large global positioning system data sets: Tests on speed-monitoring experiment in zurich. *Transportation Research Record: Journal of the Transportation Research Board*, 1935:93–100, 2005.

W. Ochieng, M. Quddus, and R. Noland. Map-matching in complex urban road networks. *Brazilian Journal of Cartography (Revista Brasileira de Cartografia)*, 55(2):1–18, 2003.

J. Pyo, D. Shin, and S. Tae-Kyung. Development of a map matching method using the multiple hypothesis technique. *IEEE Proceedings on Intelligent Transportation Systems*, pages 23–27, 2001.

M. A. Quddus, W. Y. Ochieng, and R. B. Noland. Current map-matching algorithms for transport applications: State-of-the art and future research directions. *Transportation Research Part C: Emerging Technologies*, 15(5):312–328, 2007.

N. Schuessler and K. Axhausen. Map-matching of GPS traces on high-resolution navigation networks using the multiple hypothesis technique. *Working paper*, 2009a.

N. Schuessler and K. Axhausen. Processing raw data from global positioning systems without additional information. *Transportation Research Record: Journal of the Transportation Research Board*, 2105:28–36, 2009b.

P. R. Stopher. Collecting and processing data from mobile technologies. In *8th International Conference on Survey Methods in Transport*, Annecy, France, 2008.

Q. Yang and H. N. Koutsopoulos. A microscopic traffic simulator for evaluation of dynamic traffic management systems. *Transportation Research Part C: Emerging Technologies*, 4(3):113–129, 6 1996. URL `http://www.sciencedirect.com/science/article/B6VGJ-3VWT8KB-1/2/4bfda91270dcf22f26439123b886be90`.